# A Brief Primer on Conducting Regression-Based Causal Mediation Analysis

**AUTHORS**: Yi Li[1], Kazuki Yoshida[2], Jay S. Kaufman[1], Maya B. Mathur[3]

**CORRESPONDENCE**:

Yi Li, MS.

Department of Epidemiology, Biostatistics and Occupational Health, McGill University, 2001 McGill College, Montreal, QC H3A 1G1, Canada. (Email: yi.li10@mail.mcgill.ca).

**AUTHOR AFFILIATIONS**:

1. Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, QC, Canada.

2. Division of Rheumatology, Inflammation, and Immunity, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA.

3. Quantitative Science Unit and Department of Pediatrics, Department of Medicine, Stanford University, Palo Alto, California, USA.

**WORD COUNT**: 5, 493 words

**FUNDING**: MBM received support from National Institutes of Health (NIH) grants R01CA222147, R01LM013866, UL1TR003142, P30CA124435, and P30DK116074. KY received support from the National Institute of Arthritis and Musculoskeletal and Skin Diseases (NIAMS) grants K23AR076453 and R03AR081309.

**RUNNING HEAD**: Primer on causal mediation analysis

**CONFLICTS OF INTEREST**: KY has received consulting fees from OM1, Inc. for unrelated work. The other authors declare they have no conflicts of interest.

# Supplementary Material for "A Brief Primer on Conducting Regression-Based Causal Mediation Analysis"

## 1 FORMAL DEFINITION OF CAUSAL MEDIATION ESTIMANDS AND ASSUMPTIONS

### 1.1 Effect Decomposition and Counterfactual Interpretation

To formalize the concepts of mediation introduced above, let A denote the exposure, M the mediator, Y the outcome, and C a set of all covariates. We assume the causal structure represented in the directed acyclic graph (DAG) in **Figure 1.1**. We use $a^*$ to refer to the reference level of exposure, and $a$ as the new level that is used to compare with $a^*$. (For a binary outcome, $a^*$ would simply be 1, and $a$ would be 0.) Let covariates in $C$ include all baseline A-M, A-Y and M-Y confounders and we do not need to distinguish between them.

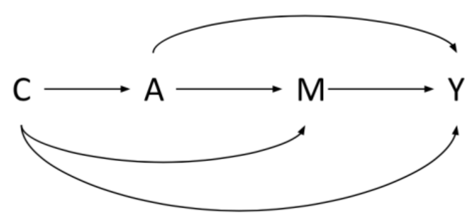


**Figure 1.1** DAG

The total effect (TE) represents the overall effect of the exposure on the outcome and is defined as

$$\text{TE} = \text{E}[\text{Y}_a \mid \mathbf{C} = \boldsymbol{c}] - \text{E}[\text{Y}_{a^*} \mid \mathbf{C} = \boldsymbol{c}],$$

where $\text{Y}_a$ represents the potential outcome had the exposure level set to $a$.

The controlled direct effect (CDE) is defined as

$$\text{CDE}(m) = \text{E}[\text{Y}_a \mid \text{M} = m, \mathbf{C} = \boldsymbol{c}] - \text{E}[\text{Y}_{a^*} \mid \text{M} = m, \mathbf{C} = \boldsymbol{c}].$$

The natural direct effect (NDE) and natural indirect effect (NIE) are defined as

$$\text{NDE} = E[Y_{aM_{a^*}} \mid \mathbf{C} = \boldsymbol{c}] - E[Y_{a^*M_{a^*}} \mid \mathbf{C} = \boldsymbol{c}],$$

$$\text{NIE} = E[Y_{aM_a} \mid \mathbf{C} = \boldsymbol{c}] - E[Y_{aM_{a^*}} \mid \mathbf{C} = \boldsymbol{c}],$$

where $Y_{aM_{a^*}}$ represents the potential outcome had the exposure level set to $a$ **and** the mediator to the level that it would have taken if the exposure level was set to $a^*$.

Please note that the NDE and NIE discussed in this paper refers to pure natural direct effect (PNDE) and total natural indirect effect (TNIE), which is a more commonly decomposition of TE. The other less common way of decomposition is total natural direct effect (TNDE) and pure natural indirect effect (PNIE). Discussions on the differences between these two decompositions can be found in methodological literature by Robins & Greenland (1992) and Pearl (2001).

## 1.2 Structural Assumptions

As noted above, certain assumptions regarding confounding are necessary to identify direct and indirect effects (Valeri & VanderWeele, 2013):

(1) There is no unmeasured exposure-outcome confounding: $Y_{am} \perp\!\!\!\perp A \mid \boldsymbol{C}$,

(2) There is no unmeasured mediator-outcome confounding: $Y_{am} \perp\!\!\!\perp M \mid A, \boldsymbol{C}$,

(3) There is no unmeasured exposure-mediator confounding: $M_a \perp\!\!\!\perp A \mid \boldsymbol{C}$,

(4) There is no exposure-induced mediator-outcome confounding: $Y_{am} \perp\!\!\!\perp M_{a^*} \mid \boldsymbol{C}$.

The first and the third assumptions hold in randomized clinical trials (RCT). The second assumption does not generally hold in RCT's because one can only randomize the exposure instead of the mediator, but it can hold in sequentially randomized trials. The fourth assumption, however, can never be guaranteed because $Y_{am}$ and $M_{a^*}$ are cross-world counterfactuals and hence can never be observed at the same time.

To identify CDE($m$), only the first two assumptions are needed. To identify NDE and NIE, all four assumptions are needed.

These four assumptions are not fully considered if using traditional approaches, given the fitted model forms. For the “difference method”, two outcome models are fitted (one regresses the outcome only on the exposure and covariates; the other regresses the outcome on the exposure, the mediator and covariates), and so users may pay more attention to controlling for exposure-outcome and mediator-outcome confounders, and neglect the third and the fourth assumptions. Similarly, for the “product method”, one mediator and one outcome models are fitted, and so users may pay more attention to controlling for exposure-outcome and exposure-mediator confounders, and neglect the second and the fourth assumptions.

In essence, the process of considering confounding control and exposure-mediator interaction is the fundamental difference between traditional approaches and the counterfactual approach. Although under extreme circumstances where all confounding is sufficiently controlled for and there is no exposure-interaction, two approaches can give the same estimates, we encourage readers to always use the counterfactual approach because of the better guarantee of causal interpretation.

## 2 R PACKAGE TUTORIAL & COMPARISON WITH TRADITIONAL METHODS

### 2.1 Causal Mediation Analysis Using *regmedint* Package

The main function in *regmedint* package is **regmedint()**. Before fitting **regmedint()**, all variables that will be used in analysis (the exposure, mediator, outcome, and confounders) should be coded as numeric. For this reason, continuous variables can be left as is, binary variables should be coded as 0/1, and any categorical variables should be converted to dummy variables, either by manual coding or a utility R package such as *fastDummies*. In our dataset, education level has four categories, so we first recoded it as three dummy variables. Categorical variables whose levels have a natural ordering, such as highest education level, could alternatively be coded as continuous, although this assumes that the variable's effects are linear.

With the analysis variables coded as numeric, we then specify the statements in **regmedint()** function:

```
regmedint_obj <- regmedint(data = dat,
                  yvar = "BSI_2020",
                  avar = "BSI_2018",
                  mvar = "SDQ_Ext_2019",
                  cvar = c("age",
                        "CERQ_positive",
                        "CERQ_negative",
                        "educ_cat2",
                        "educ_cat3",
                        "educ_cat4"),
                  a0 = 0,
                  a1 = 1,
                  c_cond = c(45, 2.11, 0.93, 0, 0, 1),
                  m_cde = 0.41,
                  mreg = "linear",
                  yreg = "linear",
                  interaction = TRUE,
                  na_omit = TRUE)
summary(regmedint_obj)
```

The name of the dataset, "dat", is specified in the argument **data**. The names of the outcome, exposure and mediator ("BSI_2020", "BSI_2018", "SDQ_Ext_2019", respectively) are specified in the arguments **yvar**, **avar**, and **mvar**.The names of the covariates are specified in **cvar**, which can take one or more variable names. We also need to specify the reference level and new level of the exposure that we want to compare using the arguments **a0** and **a1**, respectively. If the exposure is binary, then we would simply specify **a1** = 1 (i.e., exposed) and **a0** = 0 (unexposed). If the exposure is continuous, as in our example, the two levels specified define the size of the contrast in the exposure that is of interest. Here we choose 0 as the reference level (**a0**) and 1 as the new level (**a1**) to examine the effect of one-unit increase in parents' negative feelings in 2019 on their negative feelings in 2020. Please note that we do not exclude patients with exposure levels other than 0 and 1, but still use the entire datatset for analysis. The levels of **a0** and **a1** specified here only for comparing two exposure levels.

The estimated total effect, direct effect, and indirect effect can sometimes differ for individuals with different levels of the covariates (Valeri & VanderWeele, 2015; VanderWeele, 2015). Detailed discussion on when direct and indirect effects are dependent on covariates can be found in Li et al. (2022). For this reason, we must also specify what level of covariates we want to condition on, using the argument **c_cond**. A reasonable default choice is to use the sample mean levels of the covariates. For example, if we are interested in the population whose negative coping strategy, positive coping strategy and age are all at the mean levels in the study population, we would use the sample means in the argument **c_cond**, as shown in the code above. The categorical education level has three variables ("educ_cat2", "educ_cat3", "educ_cat4" in **cvar**), and we chose to obtain mediation estimates for those with master's or doctoral degrees. Thus, we set "educ_cat2", "educ_cat3", "educ_cat4" to 0, 0 and 1, respectively. Note that the values in **c_cond** should correspond to the order that variables are listed in **cvar**. We must also specify what level of mediator we are interested in. This is used to estimate conditional direct effect (CDE($m$)), but not the NDE or NIE. Again, we use the sample mean here.

The statements **mreg** and **yreg** allow us to specify the type of mediator and outcome regression models. The types of mediator models supported are "linear" and "logistic". The types of outcome models supported are "linear", "logistic", "loglinear", "poisson", "negbin" (negative-binomial), "survCox" (Cox proportional hazards), "survAFT_exp" (accelerated failure time model using an exponential distribution), and "survAFT_weibull" (accelerated failure time model using a Weibull distribution). For survival outcomes, *i.e.* if yreg = "survCox", "survAFT_exp" or "survAFT_weibull", **yvar** is the time variable, and **eventvar** is the event indicator (1 in the presence of an event, 0 if censored)." For example, if the event indicator variable in the dataset is called “event”, users need to specify **eventvar** = “event”. To include an exposure-mediator interaction term in the outcome model, the statement **interaction** should be set to TRUE, and otherwise should be set to FALSE. Given the discussion in Section 4 that including the exposure-mediator interaction captures the possibility that direction or strength of mediation differs by levels of the exposure, we would suggest setting **interaction** = TRUE by default. Please also note that including such an interaction term makes the outcome model more flexible and thus imposes fewer assumptions. On the contrary, not including the interaction term requires strong evidence that there is no interaction for all individuals in the study population.

The statement **na_omit** controls whether any missing data (coded as “NA” in R), should be removed prior to analysis. This argument is set to FALSE by default, so if there is missing data, the function will return an error message indicating that missing data are not allowed. This is designed to encourage users to check if there is missingness in the main variables of interest. If we instead specify **na_omit** = TRUE, the function will print a message indicating the number of missing values in the dataset, and a complete-case analysis will then be performed. In our example, there were six missing values (two missing exposures and four missing mediators).

Other arguments are available, but not applicable to our example. The previously mentioned argument **eventvar** is required for time-to-event outcomes. The argument **casecontrol** defaults to FALSE, indicating that the study is a cohort study, and should instead be set to TRUE if the study is a case-control study. The modeling and estimation approach discussed so far are applicable to cohort studies. For case-control studies, special modifications are needed by leaving the outcome model as is but fitting the mediator model only among controls. This requires the outcome to be rare in the population from which the controls are sampled. An alternative way (not implemented in *regmedint*) is to run a weighted mediator regression (different weights for cases and controls). Details of these two modification approaches are discussed in VanderWeele’s book (2015).

To see the output of regmedint(), we call summary(), like in lm() and glm() functions. The standard outputs of fitted mediator and outcome models will be printed out first, and then the mediation analysis results will be printed out (**Table 1**).

In **Table 1**, **cde** is the controlled direct effect, CDE(*m*), estimated for the specified mediator value. **pnde** and **tnie** are the NDE and NIE. **tnde** and **pnie** represent an alternative decomposition of total effect, which is not used as frequently as **pnde** and **tnie**; details on the subtleties between the two effect decompositions can be found in (Robins & Greenland, 1992; VanderWeele, 2013). **te** is the total effect, and **pm** is the proportion mediated.

The analysis above allowed for exposure-mediator interaction, as we recommend doing by default. If we instead wished to assume there is no such interaction, we could set the argument **interaction** = FALSE. If we do so, the estimated natural direct and indirect effects are then 0.47 and 0.02 respectively, and the proportion mediated is reduced to 0.041 (i.e, 4.1%; **Table 2**).

## 2.2 Comparison with Traditional Methods

Note that we need to exclude the rows with missing values of exposure, mediator, outcome or covariates. The dataset after removing missingness is called "dat2".

For difference method, we fit a full outcome model (including the mediator as a covariate) and a reduced outcome model (not including the mediator as a covariate):

```
y.reduc <- lm(BSI_2020 ~ BSI_2018 + age + factor(educ_cat) + CERQ_positive + CERQ_negative, data = dat2)
y.full <- lm(BSI_2020 ~ BSI_2018 + SDQ_Ext_2019 + age + factor(educ_cat) + CERQ_positive + CERQ_negative,
data = dat2)
```

As noted in the section "Comparison to traditional methods", the indirect effect is taken to be the coefficient of the exposure in the full model, and indirect effect is taken to be the difference between the coefficients of the exposure in the two models. That is, in R syntax, the difference method calculates the direct effect as coef(y.full)[2] and calculates the indirect effect as coef(y.reduc)[2] - coef(y.full)[2]. We thus obtain a direct effect of 0.47 and an indirect effect of 0.02.

For the product method, we fit the following mediator and outcome models:

```
m.fit <- lm(SDQ_Ext_2019 ~ BSI_2018 + age + factor(educ_cat), data = dat2)

y.full <- lm(BSI_2020 ~ BSI_2018 + SDQ_Ext_2019 + age + factor(educ_cat) + CERQ_positive + CERQ_negative,
data = dat2)
```

The product method takes the direct effect to be the coefficient of exposure in the full model, and the indirect effect to be the product of the coefficient of exposure in the mediator model and the coefficient of mediator in the outcome model. Namely, the product method calculates the direct effect as coef(y.full)[2] and calculates the indirect effect as coef(y.full)[3] * coef(m.fit)[2]. In this case, the product method provides the same estimates as the difference method (direct effect = 0.47 and indirect effect = 0.02). These are different from the results using causal mediation analysis if there is exposure-mediator interaction, where direct effect = 0.46 and indirect effect = 0.06 (**Table 1**).